Ann. Appl. Math.
doi: 10.4208/aam.OA-2021-0095

Vol. **38**, No. 2, pp. 183-222
May 2022

An Adaptive Gradient Method with Energy and Momentum

Hailiang Liu* and Xuping Tian

Department of Mathematics, Iowa State University, Ames, IA 50011, USA

Received 5 December 2021; Accepted (in revised version) 12 March 2022

Abstract. We introduce a novel algorithm for gradient-based optimization of stochastic objective functions. The method may be seen as a variant of SGD with momentum equipped with an adaptive learning rate automatically adjusted by an 'energy' variable. The method is simple to implement, computationally efficient, and well suited for large-scale machine learning problems. The method exhibits unconditional energy stability for any size of the base learning rate. We provide a regret bound on the convergence rate under the online convex optimization framework. We also establish the energy-dependent convergence rate of the algorithm to a stationary point in the stochastic non-convex setting. In addition, a sufficient condition is provided to guarantee a positive lower threshold for the energy variable. Our experiments demonstrate that the algorithm converges fast while generalizing better than or as well as SGD with momentum in training deep neural networks, and compares also favorably to Adam.

AMS subject classifications: 65K10, 90C15, 68Q25

Key words: Stochastic optimization, SGD, energy stability, momentum.

1 Introduction

Stochastic gradient descent (SGD) [33] is now one of the most dominant approaches for training many machine learning (ML) models including deep neural networks (DNNs) [8]. In each iteration, SGD only performs one parameter update on a mini-batch of training examples. Hence it is simple and has been proven to be efficient,

*Corresponding author.
Emails: hliu@iastate.edu (H. Liu), xupingt@iastate.edu (X. Tian)

especially for tasks on large datasets [3, 13, 43]. However, the variance of SGD can slow down the convergence after the first few training epochs; a decaying step size typically has to be applied, which is one of the major bottlenecks for the fast convergence of SGD [3,36]. In recent years, adaptive variants of SGD have emerged and shown successes for their automatic learning rate adjustment. Examples include Adagrad [7], Adadelta [45], RMSprop [41], and Adam [17]; while Adam, which may be seen as a combination of RMSprop and an exponential moving average of the first moment, stands out in this family of algorithms and stays popular on various tasks. However, training with Adam or its variants typically generalizes worse than SGD with momentum (SGDM), even when the training performance is better [43]. This explains why SGD(M) remains as a popular alternative.

AEGD (Adaptive gradient decent with energy) [21] is another gradient-based optimization algorithm that outperforms vanilla SGD. The distinct feature of AEGD is the use of an additional energy variable, which is updated together with the solution. The resulting algorithm is unconditionally energy stable (in the sense detailed in section 2) regardless of the base learning rate. Moreover, the element-wise AEGD allows for different effective learning rates for different coordinates, which has been empirically verified more effective than the global AEGD, see [21]. With AEGD the effective learning rate is the base learning rate multiplied by the energy term, against a transformed gradient.

With Adam-like adaptive gradient methods, adaptation is realized by the normalization in terms of the running average of the second order moment. While these algorithms have been successfully employed in several practical applications, they have also been observed to not converge in some other settings mainly due to the relative sensitivity in such adaptation. Indeed, counterexamples are provided in recent works [4, 24, 32] to show that RMSprop and Adam do not converge to an optimal solution in either convex or non-convex settings. In contrast, the motivation in AEGD is drawn from the perspective of dynamical systems with energy dissipation [21]. AEGD is unconditionally energy stable with guaranteed convergence in energy regardless of the size of the base learning rate and the shape of the objective functions. This explains why the method can have a rapid initial training process as well as good final generalization performance.

On the other hand, it has been long known that using momentum can help accelerate gradient descent, hence speeding up the convergence of vanilla Gradient Decent (GD) [29]. For many application tasks, momentum can also help reduce the variance in stochastic gradients [31, 35]. Using momentum has become a popular tectnique in order to gain convergence speed significantly [1, 17, 39].

With all these observations, a natural question is:

Can we take the best from both AEGD and SGDM, i.e., design an al-